63-2-5

TECHNICAL MEMORANDUM

[TM SERIES]

TM-505

COPY NUMBER _____

ASSIGNED TO _____

IMPORTANCE SAMPLING IN MONTE CARLO ANALYSIS

Charles E. Clark

24 June 1960

SYSTEM DEVELOPMENT CORPORATION
2500 COLORADO AVE.
SANTA MONICA, CALIF.

IMPORTANCE SAMPLING IN MONTE CARLO ANALYSES

ABSTRACT

Importance sampling is described as used in Monte Carlo analyses. An intuitive justification of the procedure is developed through a non-mathematical consideration of the fundamental random processes involved. The sampling procedure and its efficiency are illustrated by numerical examples.

1. Introduction. The author has consulted with operations analysts concerning the statistical problems of Monte Carlo sampling. Inevitably importance sampling is suggested, and this procedure disturbs the analyst. The difficulty is not simply that importance sampling is not understood, but that superficially it appears absurd. For example, if a Monte Carlo analysis is to evaluate the effectiveness of a weapon one of whose parameters is a reliability coefficient known to be between .50 and .75, the analyst might be told to carry out the simulation using .25 for the reliability coefficient. Such a proposal can be puzzling, and can generate resistance that is not easily overcome.

The limited understanding of importance sampling is unfortunate. The technique is easy to employ, at least in its simplest form. It can be highly efficient. When understood, it is a simple, natural procedure

which does not require professional ability in statistics. The following exposition was written for the author's clients. The discussion is intended to be an elementary presentation of fundamental statistical ideas which should be familiar to an operations analyst interested in Monte Carlo.

This paper is an expository, largely non-technical discussion of statistical sampling problems that arise in Monte Carlo analyses. Except for the appendices, no statistical knowledge is presumed beyond recognition of the nature of a probability distribution. Techniques are not elaborated. In relatively simple Monte Carlo analyses the procedures discussed can be employed adequately by the non-mathematician. In the case of an elaborate Monte Carlo, the ideas of this paper should form the basis for coordination between the operations analysts and the mathematical statistician.

The paper starts with an informal statement of what is meant by a Monte Carlo analysis. There follows a digression on stratified sampling; this digression will bring to light some important elements in the Monte Carlo analysis. Finally the discussion of importance sampling in Monte Carlo statistical analysis is presented through simple numerical illustrations. Mathematical derivations are placed in appendices.

2. Monte Carlo. This section indicates what is meant by a Monte Carlo analysis. The discussion introduces a simple example which will be used later as a numerical illustration.

Suppose that a machine starts at time zero and runs until the time of failure x. The time x is random with probability density $\lambda \exp(-\lambda x)$,

$0 \leq x < \infty$. At the time of failure the machine must be scrapped with probability p, $0 < p < 1$, but with probability $q = 1 - p$ the machine is repaired. If repaired, the machine runs from time x to $x + x'$ with x' distributed as x. Again the machine survives with probability q, and in case of survival the third failure occurs at time $x + x' + x''$ with x'' distributed as x. The process terminates when the machine is scrapped.

Suppose that we wish to know the probability that the machine will survive until time X (there may be failures before time X, but each of these failures is repaired). This probability can be computed analytically, and this computation appears below in Appendix B. Alternatively one could use the following analysis. One would draw a random number from the exponential distribution with probability density $\lambda \exp(-\lambda x)$, and this number would simulate the time to the first failure. Another random number (uniformly distributed) would determine whether the machine could be repaired. If a repair is effected, a second generation from the exponential distribution would determine the time between the first and second failures. It is obvious how the simulation would continue until the machine would be scrapped. If such a process were carried out several times, the fraction of times that the machine survived to time X would be used as an estimate of the desired probability.

If one's interest were in this problem per se, the analytic solution is much to be preferred to the statistical sampling procedure. However later in the paper we shall consider a Monte Carlo analysis of this

problem. The fact that the problem can be handled analytically will permit evaluations of the Monte Carlo analysis that would be impossible in case of a problem appropriate for Monte Carlo analysis; typically a Monte Carlo analysis is used only when an analytic solution is not obtainable. In this paper, somewhat incorrectly, an "analytic" procedure is one that does not involve statistical sampling.

The statistical sampling procedure as described above is based upon a model whose random elements are given analytically. This distinguishes the problem from a typical survey statistics problem. If one were to estimate the tobacco consumption per capita from a sample, one might consider the consumption of an individual to be a random variable. But in that case, the distribution of the random variable is unknown. One could not replace a survey of people by some desk procedure of simulating people and designating their consumptions by numbers read from a table. However, regardless of whether the sample data are obtained from a desk simulation or a field survey, the subsequent mathematical analysis of the sample data could be the same.

Some writers would distinguish between the machine-failure and tobacco-consumption problems by saying that the first can be solved by a Monte Carlo analysis, the term Monte Carlo indicating that one knows explicitly the distributions of all the random elements in the problem. In this sense the term Monte Carlo signifies that one could simulate the random process by a desk calculation which uses tables of random numbers or by a computer program which generates random numbers. With this

definition Monte Carlo does not require any distinctive mathematical analysis. The techniques of analysis were in use before the term Monte Carlo was employed. The problems to be considered in this paper are Monte Carlo problems in the sense of the above definition. The objective of the paper could be stated as that of efficiency in Monte Carlo analyses. We shall retain this definition at present, but an alternative definition will appear below.

Monte Carlo analysis, as so defined, is <u>almost</u> a general, effective procedure which enables one to solve many problems too complex for mathematical analysis. But there is one unfortunate fact. Such Monte Carlo analysis is costly. In one problem it required a high-speed computer to run 1½ hours to obtain a single sample value. At least 20 runs were required for even a small sample, and results were desired for hundreds of sets of model parameters.

There are ways to reduce the cost of such Monte Carlo analyses. Computer capabilities can be increased, and judicious adaptation of models can reduce costs. But a much easier way to reduce costs is through the employment of efficient sampling techniques. The nature and efficacy of importance sampling, one of these techniques, is the subject of this paper.

In importance sampling one considers a statistical sampling problem of the type designated above as a Monte Carlo problem. However, one does not carry out the sampling in the manner suggested by the problem. Rather a new random process is introduced in place of the original. The nature of this substitution will come to light in later sections of the paper. At present we merely remark that some writers reserve the term Monte Carlo for

a method of analysis in which one creates a random variable whose expected value is the solution of a given problem. This random variable is artificial with respect to the given problem. In the machine-failure problem one is concerned with the random variable which is 1 if a machine is scrapped prior to time X, and which is 0 if the machine survives until time X (the expected value of this random variable is the probability that a machine is scrapped prior to time X.) This random variable is given in the statement of the problem, and it is not created by the mathematician during the course of the analysis. However, in the solution constructed below by a Monte Carlo analysis, this random variable is not used. Rather the mathematician creates another random variable, whose expected value is the same as that of the given random variable, but whose expected value is cheaper to obtain by statistical sampling.

3. Stratified Sampling. The problem of section 2 could be analysed by simulating the histories of many machines and computing statistics of the outcomes. The statistician would say that data were obtained by simple random sampling. But in costly statistical analyses it is usually possible to replace simple random sampling by some more efficient procedure. Before describing such a procedure for use in Monte Carlo analyses, we shall examine some features of stratified sampling. This digression will illustrate in simple form the basic idea to be employed in importance sampling.

As a hypothetical illustrative example we suppose that a hotel wishes to estimate the mean annual expenditures by its guests in barber shops and

beauty parlors. It is known that the expenditures by women differ more widely than expenditures by men. Many men get a $2 haircut every 2 weeks at an annual cost of roughly $50; expenditures of as much as $100 or as little as $25 are found occasionally. Expenditures by women can vary from nothing to over $500. The mean expenditure for women is harder to estimate than the mean for men.

We assume that 80% of the hotel guests are men. Suppose that a sample of size 15 is to be taken (we use an absurdly small sample size to simplify the exposition.) If simple random sampling were employed, we would expect the sample to consist of 12 men (80% of 15) and 3 women. However, one might decide to sample 5 men and 10 women. Suppose the expenditures of the members of such a sample turned out to be in dollars:

Men: 50, 50, 50, 50, 100

Women: 0, 50, 100, 100, 200, 200, 200, 300, 500, 800.

It is intuitively clear that such data will lead to a more accurate estimate of the over-all average than would the expenditures of 12 men and 3 women.

To analyse these data we calculate $\bar{M}$ and $\bar{W}$, the means of the 5 male expenditures and the 10 female expenditures, respectively. The results are

$$\bar{M} = 60, \qquad \bar{W} = 245.$$

These means emphasize the fact that stratified sampling is more advantageous than simple random sampling in the present situation. We can observe that 50% of the women have expenditures greater than the mean $\bar{W}$. This reflects the fact that a minority of the women have an important influence on the mean $\bar{W}$ and on the mean when both sexes are pooled into a single distribution. It is likely that the estimates to be made would be more accurate if an even greater fraction of the sample consisted of women. However the optimum fraction is not relevant to the following discussion*.

We return to the problem of estimating the mean expenditure for all persons, male and female. If a simple random sample of size 15 had been drawn, one would divide the sum of the 15 data by 15. But this can not be done in the present instance because we have distorted the natural, simple random sampling procedure. However the analysis in the face of this distortion is obvious. Since 80% of the guests are men, we compute the following weighted mean of $\bar{M}$ and $\bar{W}$, and we obtain an estimated mean expenditure of all hotel guests to be

---

*The data suggest that the sample of women should be roughly 2.7 times as large as the sample of men. This ratio is obtained from (.2)(48.20) /(.8)(17.89) in which .2 and .8 are the fractions of women and men respectively in the population, and 48.20 and 17.89 are empirical estimates of the standard deviations of the expenditures for women and men, respectively. A justification of this result is beyond the scope of this paper. Discussions of this analysis appear in [1], [2], [3], and other discussions of stratified sampling.

$$\bar{x} = .8\bar{M} + .2\bar{W}$$

$$= (.8)(60) + (.2)(245) = 97.$$

We could estimate the sampling error in this estimate. We shall not do so because the error analysis is not needed for our purposes.

We turn next to a cruder and more cumbersome analysis of the data given above. This alternative analysis is less appealing in the barber-beauty shop problem. However, interesting analogies with Monte Carlo analysis will appear.

Let us suppose that the sample was taken among the hotel guests registered at a specific time (we ignore the fact that the statistical properties of these guests may not accurately reflect the statistical properties of all guests over a period of time.) Let us suppose that when the sample was drawn, there were 80 men and 20 women registered at the hotel. If simple random sampling had been employed, 15 of the 100 guests, without consideration of sex, would have been selected in such a manner that each guest had the probability .15 of being included in the sample. Thus a random process is visualized which would select 15 guests. Before the process would be implemented, the particular 15 selected would be uncertain, but each of the 100 guests would have the probability .15 of being selected in the sample.

This simple random sampling process was not employed. Rather the natural process was distorted. Whereas any man M_i would have the probability

$$p(M_i) = .15$$

of being included in a simple random sample, the probability was distorted to

$$p^*(M_i) = 5/80 = .0625$$

under the distorted sampling procedure which selected 5 of the 80 male guests. For any individual woman W_i the probability of being included in a simple random sample is

$$p(W_i) = .15,$$

and the probability of being included in a sample drawn by the distorted process is

$$p^*(W_i) = 10/20 = .5 .$$

Consider a particular man who was selected in the sample that was drawn. To be specific suppose that this man is the one with expenditure 100. We shall designate him by M_{100}. For analytic purposes to be revealed below, we compute for this man the weight

$$w(M_{100}) = \frac{p(M_{100})}{p^*(M_{100})} = \frac{.15}{.0625} = 2.4 .$$

The interpretation of this weight is that M_{100} would expect to appear in

simple random samples (if a large number of samples would be drawn) 2.4 times as often as in samples drawn under the distorted process. The distorted process underestimates the importance of M_{100} by the factor of 2.4. Suppose that the hotel guests numbered thousands, instead of 100, and that there were many duplicates of M_{100}. The distorted sampling process would include several duplicates of M_{100}, but in simple random sampling one would expect 2.4 times as many of such duplicates. Hence in the analysis, which will be carried out with use of formulas designed for simple random sampling, we will count M_{100} as 2.4 individuals.

Similarly, consider one of the women drawn into the sample, say W_{800}. For her we have the weight

$$w(W_{800}) = \frac{p(W_{800})}{p^*(W_{800})} = \frac{.15}{.5} = .3 .$$

If many simple random samples would be drawn, this lady would be drawn into the sample approximately 30% as often as she could expect to be chosen under the distorted process. Hence the distorted process overestimates the importance of the lady by a factor of 1/.3 = 3.33. In the analysis we should downgrade the lady's importance by counting her as .3 of a person.

We return to the numerical sample. For each person actually drawn into the sample we compute the weight. For each man the weight is 2.4 and for each woman the weight is .3. We compute the arithmetic mean of the 15 numbers in the sample, but we count each man as 2.4 men and each

woman as .3 women. The result is a new estimate of $\bar{x}$, called $\bar{x}'$, computed as

$$\bar{x}' = \frac{(2.4)(50) + \ldots + 2.4(100) + (.3)(0) + \ldots + (.3)(800)}{15} = 97.$$

Fortunately $\bar{x}' = \bar{x}$. It is possible to prove that this equality is to be anticipated. Such a proof is not presented in this paper, except for a special case found in Appendix B. Proofs are given in references [4] and [5].

The statistic $\bar{x}$ is simpler to comprehend than $\bar{x}'$. However the second statistic, or rather the basic ideas involved in the definition of $\bar{x}'$, can be employed in a wide variety of situations. In fact we can state the following general rule. As an estimator of a population expected value we could use a sample mean calculated from the elements of a simple random sample. Suppose, however, that instead of simple random sampling we use a sampling procedure in which the population elements have probabilities (or likelihoods) of inclusion within the sample which are different from the probabilities under simple random sampling. For each element x of the population from which the sample is drawn, let $p(x)$ and $p^*(x)$ be the probabilities that the element x would be drawn into the sample under simple random sampling and the alternative sampling process, respectively. Consider the weight $w(x) = p(x)/p^*(x)$. We can still use the sample mean as an estimator of the population expected value if we weight each sample value by $w(x)$. This rule will be illustrated and clarified below.

4. Importance Sampling in Monte Carlo Analysis. We are ready to discuss importance sampling. The discussion continues through the medium of trivial numerical illustrations..

Consider the exponential distribution with probability density

$$(1) \qquad p(x) = .01 \exp(-.01x), \qquad 0 \leq x < \infty .$$

We shall estimate the probability that a sample value from this distribution is less than 1. This probability is easy to obtain analytically, being $1 - \exp(-.01) = .00995$ to five decimal places. However we shall attack the problem by a Monte Carlo analysis in order to obtain a simple illustration involving sampling with distorted probability distributions.

Suppose we were to generate a simple random sample from the distribution (1). It would require a large sample to give an accurate estimate of the probability that a sample value of (1) is less than 1. This is due to the fact that approximately 1% of the sample values would be less than 1. Hence hundreds of sample values would be required before we would know that the fraction is near .01.

In order to obtain a greater proportion of sample values within the interval of importance, namely (0,1) we shall distort the sampling procedure. We introduce the distribution with probability density

$$(2) \qquad p^*(x) = \exp(-x).$$

If we sample from this distribution, which superficially has no relevance

to the problem, we shall achieve the result that a large fraction of the sample values will fall within the importance interval (0,1); the expected fraction is $1 - e^{-1} = .63$. Setting aside momentarily any question of the sanity of our operation, let us consider a sample from the distribution $p^*(x)$. Suppose that the first number generated from $p^*(x)$ were 2. Let us consider the likelihoods of generating this value 2 in both undistorted and distorted sampling. The likelihood in case of undistorted sampling is obtained from (1) as $p(2) = .01 \exp(-.02) = .0098020$, and the likelihood of drawing this same value 2 in distorted sampling is obtained from (2) as $p^*(2) = .13534$. The ratio of these likelihoods is

$$\frac{p(2)}{p^*(2)} = \frac{.0098020}{.13534} = .07$$

approximately. This implies that in undistorted sampling one can expect approximately 7% as many sample values in the interval (2, 2 + dx) as would be obtained under distorted sampling. But this means that one can sample from $p^*(x)$, count the number of sample values between 2 and 2 + dx, and multiply by .07; in this way one has an unbiased estimate of the number of sample values expected between 2 and 2 + dx under undistorted sampling (and with the same sample size.) In practice, if 2 were generated under distorted sampling, one would accept 2 not as one value but as .07 of a value.

The numbers computed above for $x = 2$ appear in Table 1. This table also contains similar results for other values of x. For example, Table 1

gives the weight 1.412 for the sample value x = 5. This implies that a sample value within the interval (5, 5 + dx) can be expected 41.2% more often with undistorted sampling than with distorted sampling. Several such weights are listed in Table 1. The weights reflect the obvious fact that small sample values are more likely to be generated from $p^*(x)$ but large values are more likely from p(x).

To illustrate the use of weighted sampling we have drawn a random sample of size 10 from $p^*(x)$. The sample values of x are listed in Table 2. In addition Table 2 gives each of the weights. Since we are estimating the probability that x is less than 1, we consider the six values of x in Table 2 that are less than 1. The value .31 is counted as .014 of an observation, .17 as .012 of an observation, etc. The sum of the weights for the six x's less than 1 is .096. Hence we count slightly less than one-tenth of an x less than 1. Since the sample size is 10, we estimate the probability that x in undistorted sampling will be less than 1 to be .096/10 = .0096. This estimate is close to the true value .00995.

The procedure has been the following. If one were to sample from p(x), approximately one out of a hundred sample values would be less than 1, and it would require a large sample to produce adequate data for an estimate of the probability that x is less than 1. We replaced p(x) by $p^*(x)$ which generates a large fraction of its sample values less than 1. We observed that any sample value from $p^*(x)$ can be weighted in such a way as to represent a number of sample values from the distribution of p(x). This number (weight) is in some cases a small fraction and in other cases much greater than 1. In the numerical illustration the distorted sampling

produced 6 of 10 sample values less than 1. But the weighting procedure led to counting each of the 6 as a small fraction of a single value when the values are to be interpreted as from the distribution of p(x). The mathematical justification of the weighting procedure and the estimate of the variance will not be made in this paper*.

5. Common Distortions of Two or More Random Processes. In Section 4 we estimated a parameter of the distribution p(x) given by (1). We did not generate a sample from this distribution; instead, our sample was from the distribution $p^*(x)$ given by (2). Let us observe that $p^*(x)$ can be regarded as a distortion of many distributions. Hence the sample of Table 2, drawn from $p^*(x)$, can be used for statistical analyses of many distributions.

To clarify this matter by a numerical illustration, we consider the distribution with probability density

$$p'(x) = .02 \exp(-.02x), \qquad 0 \leq x < \infty .$$

We shall estimate the probability that x, randomly drawn from p'(x), is less than or equal to 1. Our new problem is identical with the problem of Section 4 except that p(x) is replaced by p'(x).

---

*See [4] or [5].

We shall use the same $p^*(x)$ as a distortion of $p'(x)$. We proceed as in Section 4 and obtain Table 3 in place of Table 2. The sum of the weights in Table 3 for sample values in the interval (0,1) is .190. Dividing this sum by the sample size 10, we obtain .0190 as the estimate of the probability that x from $p'(x)$ is less than or equal to 1. This estimate can be compared with the true value .0198.

The salient feature is that two problems have been solved by use of the same sample (the first columns of Tables 2 and 3 are identical). In a serious Monte Carlo most of the computing time is used in obtaining the sample values from the distorted distribution; typically the time for statistical analysis is relatively insignificant. In our trivial example this does not happen to be true. But if we should assume that the major part of the computation consisted in the generation of the first column in Tables 2 and 3, we would conclude that we have solved two problems at the cost essentially of a single analysis.

In general, consider the probability distributions obtained by assigning a set of values to λ in $\lambda \exp(-\lambda x)$. Suppose that for each of these distributions we wish to know the probability that x is less than or equal to 1. All these problems can be solved from a single sample drawn from $p^*(x)$. If the number of values assigned to λ is large, the savings obtained from distorted sampling can be tremendous. (However, if the values of λ differ greatly among themselves, it is possible that a common distortion of all the distributions may not be efficient for every λ. It might be necessary to group the values of λ into sets, and to handle the sets separately. Such technicalities are beyond the scope of this paper.)

6. Complex Stochastic Processes. In the example of Section 4 the efficiency of the Monte Carlo analysis can be greatly increased by distorted sampling. (We say that a first sampling procedure is k times as efficient as a second procedure if the sample sizes N_1 and N_2, respectively, required for a given sampling error satisfy $N_2 = kN_1$.) Unfortunately most Monte Carlo analyses are applied to more complex stochastic processes, and the dramatic savings of Section 4 are much harder to obtain. (But the procedure of Section 5 is no less efficient.) We shall illustrate this fact by an example of a stochastic process with two random elements.

Consider the random variable y which is distributed uniformly between 0 and 200. The probability density of y is

$$P(y) = \begin{cases} 1/200, & 0 \le y \le 200, \\ 0 \text{ otherwise.} \end{cases}$$

We also use the random variable x with probability density p(x) given by (1). We assume x and y independently distributed. We shall study $z = x + y$, and we consider the estimation by Monte Carlo analysis of the probability that z is less than 1. To obtain values of z in the important interval (0,1) we shall distort the distributions of both x and y. The distortion of p(x) will be the same used above. The distortion of P(y) will be the probability density

$$P^*(y) = \begin{cases} 1, & 0 \le y \le \frac{1}{2}, \\ 1/399, & \frac{1}{2} < y \le 200, \\ 0 \text{ otherwise} \end{cases}$$

Under distorted sampling for y, i.e., with y generated from the distribution with density P^*, the weight will be $P(y)/P^*(y) = .005$ if $0 \leq y \leq \frac{1}{2}$, and $P(y)/P^*(y) = 1.995$ if $\frac{1}{2} < y \leq 200$.

Suppose that one should generate under distorted sampling $x = .4$ and $y = .4$, and hence $z = .8$. Since x and y are independently distributed, the likelihood of this pair of drawings under undistorted sampling is $p(x)P(y)$, and the likelihood under distorted sampling is $p^*(x)P^*(y)$. Hence the weight associated with the pair of values is

$$\frac{p(x)P(y)}{p^*(x)P^*(y)} .$$

This is the product of the weights associated with x and y individually. For $x = .4$, the weight is seen in Table 1 to be .0149, and for $y = .4$, the weight is given above as .005. Hence the weight associated with z determined as $.4 + .4$ is $(.0149)(.005) = .0000745$.

Suppose that another pair of drawings gave $x = .2$ and $y = .6$, and hence again $z = .8$. One easily checks that the weight associated with the pair of generations is $(.0122)(1.995) = .024339$. The important aspect of these results is that both pairs of generations produced the same z, namely $z = .8$, but the weights are different, indeed greatly different. Thus we do not have the monotonicity of the weight as a function of distance which is apparent in Table 1. Such instability of the weights can greatly reduce the efficiency of sampling distortions. This fact is proved in Appendix A.

Let us reflect on this example. The generation of a value of z requires the generation of an x and a y. Under distorted sampling the weight associated with z is the product of the weights associated with x and y. Suppose that a small value of x is drawn. Since such a small x is more likely under distorted sampling, w(x) is small. This tends to make the weight of z small. This is fortunate because our objective under distorted sampling is to get a large number of small values of z; furthermore, the large number of z's must have small weights associated with them to prevent bias in the statistical estimates.

However, this advantageous relation between x and w(x) does not necessarily produce the same relation between z and w(z). If a small x is added to a moderately large y, the sum is a z which is not small. However, the weight w(z) might be small, being the product of a very small w(x) and a value of w(y) near 1. In other words, although there may be advantageous correlations between x and w(x) as well as between y and w(y), it is an unfortunate fact that the resulting correlation between x + y and w(x)w(y) may be weak. Thus we do not have the situation in which all small values of z have small weights and all large values of z have large weights. The serious implications of this non-monotonic relation between z and w(z) may not be apparent. However, it is proved in Appendix A that such instability of the weights can greatly reduce the efficiency of the distorted sampling.

Consider a complex Monte Carlo. There will be many random variables $x_1,\ldots,x_n$, with n in some cases greater than 1,000. Instead of

the simple relation $z = x + y$, the outcome of the process z is some complex function of $x_1,\ldots,x_n$. In general, the greater n, the more difficult it is to achieve an effective correlation between this function z and the product of the n weight factors.

In complex Monte Carlo analyses one tries to introduce distorted sampling of one or more of the random variables in the process. The objective is to obtain a relatively large amount of data within intervals of importance. Furthermore these data should have small weights to compensate for their large quantity. The data which fall outside the intervals of importance should be few in number but for that reason have large weights. For complex stochastic processes it is often difficult to determine appropriate distortions.

7. A Less Unrealistic Illustration. We have described some aspects of the statistical sampling problem in Monte Carlo analysis. These discussions will be summarized through the medium of a numerical example which is intended to bridge the gap between formalism and realistic application. The discussion is based upon mathematical results which are deferred to the appendices.

We shall study the process described above in which the running time between failures of a machine is generated from the distribution with probability density $\lambda \exp(-\lambda x)$, $0 \leq x < \infty$; at each time of failure the machine dies (is scrapped) with probability p but is repaired with probability $q = 1 - p$; in case of repair an additional running time is generated from the same exponential distribution; the process continues until death is

generated at a time of failure. Let us suppose that the basic problem is to determine the 1% quantile of the distribution of times to death. In other words we wish a lower tolerance limit for this time to death so that with 99% confidence one can assume that a machine's life will exceed this tolerance limit. This 1% quantile, which we shall denote by X, can be computed analytically, and this is done in Appendix B. For this reason our example is simpler than most Monte Carlo simulations. But the simplicity will permit analytic evaluations which are impossible if a simulation is complex.

The running times and death or survival at each failure could be simulated. The histories of several machines could be generated, and the time of death recorded for each history. From the record of these empirical times of death, one could estimate the 1% quantile X. Such an estimate would have a large relative error unless the sample size were very large. This is due to the fact that very few of the empirical data would be within the interval of importance (0,X).

To obtain a greater fraction of the empirical results within the importance interval (0,X), we can distort the Monte Carlo process. We can replace the distribution of running times with one having a smaller expected time between failures. Furthermore we can increase the probability of death at each failure. We note that the expected value of the random variable with probability density $\lambda \exp(-\lambda x)$ is $1/\lambda$ (indeed, the integral from 0 to ∞ of $\lambda x \exp(-\lambda x)$ is $1/\lambda$.) Hence if the exponential distribution with parameter λ is replaced by the exponential distribution

with parameter λ^* with $\lambda < \lambda^*$, the expected time between failures is reduced from $1/\lambda$ to $1/\lambda^*$. In addition we can replace the probability of death p by $p^* > p$.

Thus one would hope to employ importance sampling advantageously by increasing λ and p to λ^* and p^*. But appropriate values for λ^* and p^* are not immediately obvious. Should both or only one of the parameters be distorted? How great should the distortions be? Before discussing how one might resolve these questions, we shall indicate the optimum distortions in a special case.

To particularize the discussion we shall use $\lambda = 1$ and $p = \frac{1}{2}$. For several pairs of values of λ^* and $q^* = 1 - p^*$ we have computed the efficiency of the distorted sampling relative to undistorted sampling. These relative efficiencies appear in Table 4. For example Table 4 gives .00538 for the relative efficiency in case $\lambda^* = 80$ and $q^* = .005$. This means that if the sampling error is preassigned, and if a sample of size N is required under undistorted sampling to keep the error of estimate within the given limit of error, a sample size of .00538N would be adequate to attain the same accuracy if the parameters are distorted to $\lambda^* = 80$ and $q^* = .005$.

We shall present a mathematical analysis of this numerical example in Appendix B. But first we conclude the non-mathematical part of the exposition with some general remarks. In a real problem one cannot construct Table 4; if one has enough information to construct such a table, it is likely that one could solve the problem analytically. Hence one

must obtain good distortions of the model parameters partly by guesswork. Such guessing is not a simple matter. Table 4 reveals that some distortions would be disastrous.

In the face of our illustrative problem one might reason as follows. One cannot get a large fraction of small times to death x merely by increasing p. This is due to the fact that the first time to failure is generated before the probability p comes into play, and the first time to death already exceeds X in a large fraction of the histories. Hence it is likely that large sampling savings will require a distortion of λ. It would appear dangerous to rely solely on a distortion of p.

Without further insight into the times that would be generated under undistorted sampling, one could not do much better than the following. We note that it could be disastrous to use too large values of λ^* and p^*; indeed, Table 4 indicates that $\lambda^* = 1$ and $q^* = .0005$ would be bad, and we remark, without proof, that similar unfortunate results are obtained if λ^* is increased beyond the range in Table 4. Hence one might generate two small samples of histories with small distortions to say $\lambda^* = 2$ and $q^* = .4$ in one sample and $\lambda^* = 3$ and $q^* = .3$ in the other. Estimates of the sampling errors in the two samples could suggest trial values of the parameters in a third sample. This timid, tentative, probing procedure has serious disadvantages. In the first place, the sampling errors in the estimates of the sampling errors would be great with small samples, and the empirical results might mislead one to believe, until further data were available, that a distortion of λ to 2 is better than a distortion to 3. More seriously, the analysis would be completed (with

small savings) before one used parameters anywhere near the optimum. In some cases these difficulties cannot be circumvented.

If one can anticipate the results that would be obtained under simple random sampling, one can act more effectively. Suppose one had reason to believe that in undistorted sampling the expected value of the time to death is near 2 (it is 2) and that the 1% quantile is near .020 (it is .020). Then one would know that in undistorted sampling much of the time-to-death data would be near 2, whereas we would like data near .020. This suggests a distortion of the times to death which decreases the expected value of the distances by a factor of roughly 100. Use of $\lambda^* = 100$ would effect such a distortion. Hence for the sake of simplicity one could leave p undistorted and use $\lambda^* = 50$ or $\lambda^* = 25$ depending upon how timid one is in the face of the fact that it is worse, usually, to overestimate than to underestimate the optimum distortion.

It is possible to devise less elementary procedures for arriving at a good estimate of an optimum distortion (but to the author's knowledge, current literature does not indicate any simple, generally applicable procedure for estimating accurately an optimum distortion.) In general, elementary considerations often cannot be sharper than the above thoughts.

These thoughts lead to the suggestion that in the hands of a mathematical amateur, importance sampling can produce significant but moderate savings in computing time. If a Monte Carlo is so large that high computing costs are involved, it is likely that profit would result from professional mathematical assistance in the design of the statistical sampling procedures.

8. Estimates of Expected Values. The entire paper has been limited to the discussion of one problem, the estimation of the probability that the output of a Monte Carlo is less than X, where X is the 1% quantile or some other quantile with a small percentage. This probability is an expected value (of the random variable which is 1 if the time to death is less than X, and is 0 otherwise.) In general, Monte Carlo analysis involves the estimation of expected values. However the different information requirements arise in the estimations of different expected values. Hence technical procedures vary from problem to problem.

Consider, for example, the numerical example discussed in Section 7. One might wish to know the expected value of the distribution of times to death. For an estimate of this expected value, the distortions used above would be bad. The large times of death are more important than the small ones in the sense that an efficient sample should contain more large times than would be generated in simple random sampling. One would in this case decrease λ and p. We shall not discuss this problem. Our purpose at this point is to warn the reader not to assume that all Monte Carlo statistical analysis are completely similar in details to the illustrations used in this paper.

APPENDIX A

In Section 6 concerned with complex stochastic processes, we considered a random process with two random elements x and y. The output of the process $z = x + y$ was such that different pairs of x and y could produce the same value of z. In addition, under the distorted sampling employed, the weights associated with the different pairs could be different. Hence for fixed z the weights vary. It was stated that this variability of the weights can greatly reduce the efficiency of the distorted sampling. This result will follow from the sampling error formula derived in Appendix A.

We have a random process, and we wish to estimate the probability that the outcome of the process is within some interval I; in the problems considered in this paper the interval I was in the form (0,X). A sample is drawn with distorted distributions, and the empirical data consist of a sequence of "distorted" outputs x_i^*, $i = 1,\ldots,N$, and the corresponding weights w_i; the outputs are sample values of a random variable. The estimate of the probability that the output (in undistorted sampling) is within I is

$$t = N^{-1} \sum_{1}^{n} w_i, \tag{3}$$

where we assume that the x_i^* have been so arranged that x_i^* is within I if and only if $i \leq n$. (The statistic t is simply the number of sample

values within I divided by the sample size, each sample value being counted w times.) We shall derive the sampling variance of t.

Let p^* be the probability that an output x^* will be within I under distorted sampling. Then n is a binomially distributed integer with N and p^* as the binomial parameters. Hence the probability of n is

$$P(n) = \binom{N}{n} p^{*n} q^{*N-n}. \tag{4}$$

The expected value of t is the sum over n of the expected value given n multiplied by the probability of n; in symbols

$$E(t) = \sum_0^N E(t|n)P(n). \tag{5}$$

Let $E(w^i|x^*\epsilon I)$, $i = 1,2$, be the conditional expected value of w^i under the condition that the corresponding distorted output x^* (i.e., the x^* which has w as its weight) is in I. Since each w_i which appears in (3) corresponds to an x^* within I, we have

$$E(t|n) = N^{-1}nE(w|x^*\epsilon I).$$

This, (4) and (5) give

$$E(t) = N^{-1}E(w|x^*\epsilon I) \sum_0^N n\binom{N}{n} p^{*n} q^{*N-n}.$$

Since the sum appearing in this expression is the expected value of n,

and since this expected value is Np^* from the well-known theory of the binomial distribution, we obtain

$$E(t) = p^* E(w^i | x^* \epsilon I). \tag{6}$$

In the calculation of $E(t^2)$ we shall encounter

$$\sum_0^N n^2 \binom{N}{n} p^{*n} q^{*N-n}$$

which is the expected value of the square of the binomially distributed variable n, which is the variance plus the square of the expected value of n, which is well-known to be

$$Np^*q^* + (Np^*)^2. \tag{7}$$

Using this result we calculate that

$$E(t^2) = \sum_0^N E(t^2|n)P(n)$$

$$= \sum_{n=0}^N N^{-2} E\left[\sum_{i=1}^n w_i^2 + 2 \sum_{1 \le i < j \le n} w_i w_j \right] \binom{N}{n} p^{*n} q^{*N-n}$$

$$= N^{-2} \sum_0^N \left\{ nE(w^2 | x^* \epsilon I) + n(n-1)[E(w|x^* \epsilon I)]^2 \right\} \binom{N}{n} p^{*n} q^{*N-n}$$

(because of the independence of w_i and w_j, $i \neq j$, we have $E(w_i w_j)$ $= [E(w)]^2$ for all $n(n-1)/2$ pairs.) We replace $n(n-1)$ by $n^2 - n$ and use (7) to obtain

$$E(t^2) = N^{-2}[E(w^2|x^*\epsilon I)Np^* + [E(w|x^*\epsilon I)]^2[Np^*q^* + (Np^*)^2 - Np^*]]$$

$$= N^{-1}p^*E(w^2|x^*\epsilon I) + (p^{*2} - N^{-1}p^{*2})[E(w|x^*\epsilon I)]^2$$

because $1 - q^* = p^*$.

The variance of t is

$$V(t) = E(t^2) - [E(t)]^2$$

which with use of (6) reduces to

$$V(t) = N^{-1}[p^*E(w^2|x^*\epsilon I) - p^{*2}[E(w|x^*\epsilon I)]^2]$$

$$= N^{-1}[p^*V(w|x^*\epsilon I) + p^*q^*[E(w|x^*\epsilon I)]^2]$$

because $E(w^2) = V(w) + [E(w)]^2$ and $p^* - p^{*2} = p^*q^*$.

This result might lead one to conclude that the sampling error can be reduced by making p^* small, i.e., by use of a distortion which produces a small amount of data within the interval of importance I. However such a conclusion would be fallacious. If p^* were small, the values of w, given $x^*\epsilon I$, would be large. The increase in $E(w|x^*\epsilon I)$ and $V(w|x^*\epsilon I)$ would

greatly outweigh the decrease in p^*. We shall not pause to develop this point. On the other hand, p^* and q^* can never exceed 1. Hence small values of the expected value and variance of w, given $x^* \epsilon I$, will produce a small sampling error. We recall that a large expected number of distorted times x^* within I must imply small $E(w|x^* \epsilon I)$. This in turn implies that most values of w, given $x^* \epsilon I$, should be small, and hence the variance of w, given $x^* \epsilon I$, should be small.

APPENDIX B

Appendix B contains a mathematical analysis of the illustrative random process used above. The time between failures is distributed with probability density

$$\lambda \exp(-\lambda x). \tag{8}$$

At each failure death occurs with probability p, but repair is effected with probability $q = 1 - p$. When death occurs the process terminates, but a repair is followed by another time to failure. We shall compute the distribution of times to death. But most of Appendix B is taken up with computation of weights and sampling error under importance sampling. The error to be studied is the sampling error in a sample estimate of the probability that a time to death does not exceed X, where X is any real number. An interesting value for X is the 1% quantile of the times to death.

As an auxiliary formula we shall derive the probability density of $x = x_1 + \ldots + x_n$ where the x_i are independently distributed with probability density (8). We shall prove that this probability density is

$$P(x|n) = \lambda^n x^{n-1} e^{-\lambda x}/(n - 1)! \quad . \tag{9}$$

For $n = 1$ the relation is obvious. To verify the general case by induction we write

$$P(x|n) = P(x_1 + \ldots + x_n) = \int_0^x \lambda^{n-1} t^{n-2} e^{-\lambda t} [1/(n - 2)!] \lambda e^{-\lambda(x-t)} dt$$

which reduces to (9).

Let n be the number of failures prior to death, and let x be the time to death. The probability that death occurs at the end of the n^{th} time between failures is

$$P(n) = q^{n-1}p. \tag{10}$$

The probability density of x given n is (9). Hence the probability density of x is

$$P(x) = \sum_{1}^{\infty} P(x|n)P(n)$$

which reduces to

$$P(x) = p\lambda \exp(-p\lambda x). \tag{11}$$

Thus as stated in Section 2 on Monte Carlo, our illustrative problem can be solved analytically. If $\lambda = 1$ and $p = \frac{1}{2}$, the 1% quantile of x, denoted by X, is obtained from

$$\int_{0}^{X} \tfrac{1}{2} \exp(-\tfrac{1}{2}t)\, dt = .01,$$

and this gives $X = -2 \log .99 = .020$ as stated above. However, we have introduced a Monte Carlo analysis of this problem because of the possibility of the following mathematical evaluation of the importance sampling

procedure. Such a mathematical evaluation would be impossible in case of a complex Monte Carlo.

We assume that importance sampling is employed in which λ and p are distorted to λ^* and p^*. If a distorted history leads to death at the time of the n^{th} failure, and if x_i^*, $i = 1,\ldots,n$, are the times between failures, the weight is the product

$$\frac{\lambda \exp(-\lambda x_1^*)}{\lambda^* \exp(-\lambda^* x_1^*)} \frac{q}{q^*} \frac{\lambda \exp(-\lambda x_2^*)}{\lambda^* \exp(-\lambda^* x_2^*)} \frac{q}{q^*} \cdots \frac{\lambda \exp(-\lambda x_n^*)}{\lambda^* \exp(-\lambda^* x_n^*)} \frac{p}{p^*}$$

which we write

$$w(x^*,n) = \left(\frac{q}{q^*}\right)^{n-1} \frac{p}{p^*} e^{(\lambda^*-\lambda)x^*} \tag{12}$$

where

$$x^* = x_1^* + \ldots + x_n^*$$

is the (distorted) time to death.

Our problem is to estimate the probability (in undistorted sampling) that a time to death x is less than or equal to X. If N is the sample size, and if the index i indicates the i^{th} sample history, our estimator is

$$N^{-1} \sum_{i=1}^{N} C_X(x_i^*) w(x_i^*, n_i)$$

where $C_X(t) = 1$ if $t \leq X$ and $= 0$ if $t > X$. This means that we count the histories for which $x^* \leq X$ and divide by the number of histories in the sample, the history x_i^* being counted as $w(x_i^*, n_i)$ histories. We shall study the random variable

$$t = C_X(x^*)w(x^*,n),$$

because the variance of the sampling statistic is $N^{-2}V(t)$, where V denotes variance.

For the expected value of t we have

$$E(t) = \int_0^\infty \sum_{n=1}^\infty C_X(x^*)w(x^*,n)P(x^*|n)P(n)dx^*.$$

The factor $C_X(x^*)$ can be deleted if we integrate over (0,X) rather than $(0,\infty)$. Hence using (9), (10), and (12) we obtain

$$E(t) = \int_0^X \sum_{n=1}^\infty \left(\frac{q}{q^*}\right)^{n-1} \frac{p}{p^*}\, e^{(\lambda^*-\lambda)x^*} \frac{\lambda^{*n} x^{*n-1} \exp(-\lambda^* x^*)}{(n-1)!}\, q^{*n-1} p^* dx^*$$

$$= p\lambda \int_0^X e^{-p\lambda x^*} dx^*$$

$$= 1 - \exp(-p\lambda X).$$

The reader can check using (11) that E(t) is the probability that $x \leq X$. This justifies in this particular case the assumption made without proof that the weighted output of distorted sampling produces an unbiased estimate.

We obtain $E(t^2)$ by replacing w by w^2 in the initial expression for $E(t)$. Reductions similar to those above give

$$E(t^2) = \frac{p^2 q^* \lambda^2}{p^*(q^*\lambda^{*2} - 2q^*\lambda\lambda^* + q^2\lambda^2} \left\{ -1 + \exp\left[X(\lambda^* - 2\lambda + \frac{q^2\lambda^2}{q^*\lambda^*} \right] \right\}.$$

The variance of t is obtained as $E(t^2) - [E(t)]^2$, and Table 4 is obtained from the specialization of $V(t)$ with $\lambda = 1$ and $p = \frac{1}{2}$. Table 4 is this $V(t)$ divided by the value of this variance in undistorted sampling, that is, when $\lambda^* = \lambda$ and $p^* = p$.

REFERENCES

1. W. G. Cochran, Sampling Techniques, John Wiley and Sons, New York, 1953.

2. W. E. Deming, Some Theory of Sampling, John Wiley and Sons, New York, 1950.

3. M. H. Hansen, et al, Sample Survey; Methods and Theory, Vols. I and II, John Wiley and Sons, New York, 1953.

4. H. A. Meyer, ed., Symposium on Monte Carlo Methods, John Wiley and Sons, New York, 1956, Chapters 8, 9, and 15.

5. H. Kahn, Applications of Monte Carlo, The RAND Corporation, Santa Monica, California, 1956.

TABLE 1

Weights related to the use of $p^*(x)$ as a distortion of $p(x)$

x	p(x)	$p^*(x)$	$w(x) = p(x)/p^*(x)$
.1	.0099900	.90484	.0110
.2	.0099800	.81873	.0122
.3	.0099700	.74082	.0135
.4	.0099600	.67032	.0149
.5	.0099501	.60653	.0164
.6	.0099402	.54881	.0181
.7	.0099302	.4[illegible]659	.0200
.8	.0099203	.44933	.0221
.9	.0099104	.4[illegible]657	.0244
1.0	.0099005	.36788	.0269
2.0	.0098020	.13534	.0724
3.0	.0097049	.049787	.1949
4.0	.0096090	.018316	.5246
5.0	.0095143	.0067379	1.412
6.0	.0094176	.0024788	3.799
7.0	.0093239	.00091188	10.22
8.0	.0092311	.00033546	27.52
9.0	.0091393	.00012341	74.56
10.0	.0090484	.000045400	199.3

TABLE 2

Sample from $p^*(x)$ as a distortion of $p(x)$

Sample x from $p^*(x)$	$p(x)$	$p^*(x)$	$w(x) = p(x)/p^*(x)$
2.71	.009733	.06654	.146
.31	.009969	.7334	.014
.17	.009983	.8437	.012
.02	.009998	.9802	.010
.59	.009941	.5543	.018
.54	.009946	.5828	.017
4.15	.009594	.01576	.609
.91	.009909	.4025	.025
2.72	.009732	.06588	.148
1.15	.009886	.3166	.031

$$\sum_{x<1} w(x) = .014 + .012 + .010 + .018 + .017 + .025$$

$$= .096$$

$$(10)^{-1} \sum w(x) = .0096$$

TABLE 3

Sample from $p^*(x)$ as a distortion of $p'(x)$

Sample x from $p^*(x)$	$p'(x)$	$p^*(x)$	$w(x) = p'(x)/p^*(x)$
2.71	.01894	.06654	.285
.31	.01988	.7334	.028
.17	.01993	.8437	.024
.02	.01999	.9802	.020
.59	.01976	.5543	.036
.54	.01978	.5828	.034
4.15	.01841	.01576	1.17
.91	.01964	.4025	.049
2.72	.01894	.06588	.288
1.15	.01954	.3166	.062

$$\sum_{x<1} w(x) = .027 + .024 + .020 + .036 + .034 + .049$$

$$= .190$$

$$(10)^{-1} \sum w(x) = .0190$$

TABLE 4

Sample error in distorted sampling divided by the error in undistorted sampling.

λ^* \ q^*		.5	.3	.1	.01	.005	.001	.0005
1	1.00000	.70904	.55679	.54068	.64145	.84770	14.58096	1087.46222
5	.19739	.13826	.10571	.10018	.10013	.10518	1.78017	3.33124
10	.09908	.06796	.05075	.04552	.04653	.04762	.06128	.08494
20	.05055	.03327	.02370	.02197	.02100	.02126	.02470	.02989
40	.02770	.01694	.01097	.00988	.00919	.00923	.01027	.01143
60	.02167	.01264	.00763	.00671	.00610	.00610	.00668	.00754
80	.02032	.01167	.00687	.00598	.00540	.00538	.00580	.00643
100	.02132	.01239	.00743	.00652	.00590	.00587	.00622	.00695
120	.02409	.01436	.00896	.00798	.00729	.00725	.00758	.00807
140	.02854	.01754	.01145	.01032	.00949	.00949	.00981	.01030
160	.03492	.02210	.01499	.01368	.01278	.01270	.01303	.01353
180	.04361	.02833	.01982	.01825	.01718	.01709	.01743	.01800

CEC:rs

20 June 1960

UNCLASSIFIED

System Development Corporation,
Santa Monica, California
IMPORTANCE SAMPLING IN MONTE CARLO
ANALYSIS.
Scientific rept., TM-505, by
C. E. Clark. 24 June 1960, 41p.,
4 tables.

Unclassified report

DESCRIPTORS: Monte Carlo Method.
Statistical Distribution.

Develops a intuitive justification
of importance sampling in Monte
Carlo analyses through a non-mathematical

UNCLASSIFIED

UNCLASSIFIED

consideration of fundamental random
processes. Numerical examples
illustrate its efficiency.

UNCLASSIFIED